Text Mining for User Query Analysis

A 5-Step Method for Cultural Heritage Institutions

Anne Chardonnens, Simon Hengchen

Université libre de Bruxelles, Belgium
{anchardo, shengche}@ulb.ac.be

Abstract

The recent development of Web Analytics offers new perspectives to libraries, archives and museums to improve their knowledge of user needs and behaviours. In order to dive into the mind of their end users, institutions can explore queries from a digital catalogue. However, a manual exploration demands a major time commitment and only leads to limited results. This paper explores how text mining techniques can help automate the analysis of large volumes of log files. A 5-step methodology including clustering is illustrated by a case study from the State Archives of Belgium.

Keywords: user query; logs analysis; information retrieval; text mining; cultural heritage

In: M. Gäde/V. Trkulja/V. Petras (Eds.): Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017), Berlin, 13th—15th March 2017. Glückstadt: Verlag Werner Hülsbusch, pp. 177—189.

1 Introduction

In a context of budget cuts[1] and an increasing impulse to reach wider and more diverse audiences, cultural heritage institutions require a better understanding of user needs. Moreover, as expectations regarding content and services have grown with the evolution of the Web, it is essential for the institutions to understand the changing behaviour of users (Showers, 2015).

Beyond methods institutions developed to monitor the way collections are used *in situ*, from rare medieval maps to the copies of Anna Karenina, they have to imagine ways to monitor online activities. Web Analytics offer a new quantitative method to understand user behaviour. This unbiased observational data represents a unique opportunity to dive more precisely into the mind of end users through requests they made into the catalogue search engines (Grimes, Tang & Russel, 2007). From this approach, libraries, archives and museums can gain three kinds of insights.

First, by gathering statistical data, they can identify the types of content which turn out to be the most popular among theirs users. This information can be used as priority indicators regarding acquisition, cataloguing and digitisation of collections in order to meet user demands better. Secondly, user queries can be explored to better understand end users' behaviours and improve users' experience. Thirdly, they can discover the way users formulate what they seek and thus adapt the collection's metadata.

It should be noted that manual analysis involves a major time commitment for limited results. Indeed, it is difficult to obtain reliable numbers on the popularity of a topic, considering changes in the ways of writing queries. For example, it would be very time consuming to identify and reconcile manually every query related to *registres paroissiaux* (parish registers) with variant spellings (singular, plural, with space or separated by a "+", etc.). This is the kind of task that can be semi-automatically performed via text mining techniques, leading to potentially more precise information.

This paper explores the application of text mining techniques for user query analysis. With the help of a concrete case study from the State Archives of Belgium, we illustrate the possibilities but also the limits of using

1 This was, for example, the case with the European Commission that not longer provides funding for the metadata creation since 2008 (van Hooland, Vandooren & Méndez Rodríguez, 2011).

text mining techniques in a cultural heritage setting. The article is constructed as follows. After this introduction, the second section presents an overview of existing works on user queries. The core of the paper consists of a 5-step method including clustering to analyse user queries, which is illustrated by our case study. The paper ends with results, discussion and future work.

2 Related works

Studying user queries and access paths to digital content in a cultural heritage context involves concentrating the attention on different research areas. At present, the way users search in digital catalogues is influenced by research conducted in other contexts. A well-known example is the Google search environment, which initiated the concept of the "Googlized" library patron (Woods, 2010).

Studies based on search engines are numerous and characterised by very different angles of approach: user information-seeking behavior, analysis of failed queries, domain knowledge or multilingual issues. Moreover, this field of research evolves rapidly, in parallel with the sophistication of search engines and the rise of new computational methods. For example, Grimes, Tang and Russell (2007) reviewed three sources of data that can be used to improve performance on the goals of a search engine: the field study, the usability study data and the raw query log. While noting the rare opportunity to gather information without disturbing the user, they underline the fact that the logs are not sufficient to measure the why, only the how and the what. Three years later, Kathuria, Jansen, Hafernik and Spink (2010) published a paper aiming to automatically classify the different user intents behind web queries.

In the cultural field, different methods have been used to investigate user search experience. To plan an academic library website redevelopment, Chase, Trapasso and Tolliver (2016) conducted a usability test. Though their study leads to actionable findings, such as a need for more support and information literacy instruction for students, it does not give much knowledge about the content of the queries themselves. More generally, Zavalina and Vassilieva (2014) observed that, among all the studies published by large scale digital libraries they reviewed, only a minority examines the content of

the user search queries. Ceccarelli, Gordea, Lucchese, Nardini and Tolomei (2011) analysed query logs to enhance the usability of the Europeana Portal and to develop assistance functionalities such as a query recommender system. Dijkshoorn et al. (2014) used log files from the Rijksmuseum to combine user queries with external vocabularies published as Linked Data, in the attempt to diversify search results. However, in both cases, no text mining methods appear to have been used on user queries.

By contrast, in the context of her analysis performed on the IMLS Digital Collection Registry transaction log dataset, Zavalina (2007) described how they first processed the queries (truncating plural, grouping together correct and misspelled versions of the same word, excluding stopwords such as prepositions, etc.) and then worked with a controlled vocabulary. However, the corpus contained less than 1 000 queries and apparently the whole process was done manually, including the extraction of all query strings. This last example illustrates the potential of text mining in this context: using a script to semi-automatically process the data could save time and be applied to large datasets. The workflow reported in this article seeks to bridge this gap by describing in detail each step and illustrating its utility.

3 Method

The present method aims at splitting the whole analysis process in five concrete steps. Since cultural heritage institutions have limited human resources at their disposal and are sometimes forced to outsource some technical tasks, special attention has been paid to provide open source solutions,[2] adaptable to various contexts. Regardless of the software architecture behind the search engine of a digital catalogue, our method allows every potential user to retrieve and exploit user queries in a semi-automated manner.

2 The code is freely available online: https://github.com/anchardo/PGCC.

3.1 Collecting

The first step of the method is to actually collect the user-entered data. In the case of Web Analytics, three different possibilities exist to collect data: web log files; query parameters and JavaScript functions. Web log files contain raw data recorded automatically by web servers; query parameters composing URLs can be provided to analytics tools such as Google Analytics, which will retrieve and store them automatically; JavaScript functions can be used to capture and store the queried terms if they do not appear in the URL. Regardless of the method chosen, once the data has been collected and stored, it becomes processable.

3.2 Parsing

Within the context of Web Analytics, parsing is the task of reading through an URL and selecting, according to predetermined rules, the terms of the query and, when available, the advanced search parameters. This task, requiring possession of web log files, proves to be essential when the aim is to carry out an in-depth analysis.

Since all websites are different, as well as the structure of the corresponding URLs, the parsing method should be adapted for each unique website. Nonetheless, what can sound like a daunting task is made easy by the use of regular expressions (the so called "regex")[3], which can be customised to extract the information needed from any URL.

3.3 Grouping

Data analysis tools provide powerful functionalities to group and aggregate data, which will be useful in such a context. Thus, queried terms entered only once during a visit may still appear several times within the next URLs (for example when the visitor consults the results pages). Counting all these occurrences indifferently would skew study results such as the most searched terms. Grouping functionalities offers the possibility of applying an approximate but consistent method consisting of keeping each query at the

3 Regex is a text pattern following a specific syntax, which helps to find any string of characters matching that pattern within a text (Goyvaerts & Levithan, 2009).

most once per visit or per visitor, which helps to have a bird's eye view of the most popular search queries.

3.4 Cleaning

At this stage, user queries have been extracted and stored in tabular files. Before being analysed, they still need to be "cleaned". The goal is to be able to associate similar string of characters despite superficial differences (Manning, Raghavan & Schütze, 2008), otherwise statistics could be skewed.[4]

That step can consist of various operations: to trim whitespaces – which means removing unnecessary spaces from the left-hand and right-hand sides of strings –, to convert all characters to lowercase or to replace special characters. Another process is tokenisation, which leads to chop each of the words composing a query, if there is a need to analyse them in an isolated manner. Finally, one could want to go further focussing on grammatical differences via stemming and lemmatisation, to reduce various forms of a word to a common base form (ibid.). Types of operations carried out during that step depend mostly on the data set and the needs of the analysis. The most important is to be consistent, i.e. to pre-process the whole data set the same way.

3.5 Clustering

This last step is designed to tackle different issues. First, user-entered data is always prone to errors. Secondly, the spelling of proper nouns is known to have evolved throughout the ages. Thirdly, the cleaning process only groups quasi perfect matches. It is therefore worthwhile to implement an additional function which seeks to cluster similar-yet-different terms. That task, called clustering, can be completed by various algorithms. The choice of the algorithm has to be made according to the data set particularities.

4 If user queries were not pre-processed, this could, for example, lead *Jan van Nijlen* and *Jan Van Nijlen* (capital "v") to be considered as different queries.

4 Case study

The selection of our case study was guided by these three steps: find an institution willing to share its log files; obtain a consistent and large data set to test text mining techniques under realistic conditions; study user queries in a multilingual context.

Our active participation in a research project involving three national cultural heritage institutions of Belgium[5] facilitated our access to raw data. Our final choice fell on the State Archives of Belgium since they also fulfil the two other conditions, with a four-language online catalogue and more than 175 000 visits per month.[6] Moreover, until now, no research had been conducted by the institution on user queries, despite it being an important issue.

The online presence of the State Archives consists of an informational website and a digital catalogue[7] called *Search*, whose user queries will be extracted as the raw data of our study. *Search* provides access to the database of the State Archives collection, which gathers archival heritage from the National Archives in Brussels and 18 repositories throughout the country. The scope of our study will focus on the main search engine, which is supplemented by two other devices enabling search by person or by producer.

4.1 Collecting

Among the three main possibilities to collect user queries (web log files, query parameters and JavaScript), we choose the first one, which allows us to gather additional information.

Search, the digital catalogue from the State Archives of Belgium, is tracked via a Piwik instance.[8] A web server automatically records log files containing raw data such as the web pages users request or the IP addresses

5 The Maddlain project, a research project with the aim of modernising digital access to the collections of the Royal Library of Belgium, the States Archives of Belgium and the CegeSoma: http://www.maddlain.iminds.be/en/.

6 Numbers based on the average monthly visits between January and June 2016, which amounts to 176 510 visits, realised by an average of 44 795 unique visitors.

7 http://search.arch.be/en

8 The Piwik integration has been carried out by imec, a research institute linked to the University of Antwerp and the Gent University.

they are requested from. It has to be noted that those files are enormous[9] and filled with data of no interest for such a study. In order to tackle performance issues due to that large amount of raw data and to avoid being forced to use more complex infrastructures, we used two strategies to reduce the amount of data to store locally. First of all, we reduced the survey period to six months (from 1 January to 1 July 2016). Second of all, we used a text filter[10] to keep only URLs containing user queries.[11]

4.2 Parsing

In order to handle our data set and apply advanced processing methods, we selected an open-source interactive programming environment: the Jupyter notebook.[12] Within that environment, we wrote a script in Python. Via three rather brief functions,[13] the queries written by users can be extracted,[14] as well as advanced parameters such as language preferences or filters based on time period or location of deposits.

4.3 Grouping

Within the Jupyter notebook, we used Pandas – an open source Python package – and more precisely its "group by" functionality to keep only one occurrence of each keyword(s) entered by a user during their visit.

In order to do so, we created a DataFrame containing three columns, based on informations parsed from the URLs: the ID of the visit, the queried

9 The intensive usage of the site (approximately 300,000 pageviews/day, mainly pages containing digitised genealogical sources) produces more than 2 GB of log files per month.

10 A regex inserted within the SQL LIKE operator.

11 Less than 1% of the URLs that correspond to the 300 000 pageviews/day contain search terms entered into the main search engine (the majority of users arrive at digital sources by means of links rather than by making use of search functionalities).

12 http://jupyter.org/

13 For the sake of brevity, the code is not presented here, but it is available online: https://github.com/anchardo/PGCC.

14 Here is an example of an URL containing the specific search terms ("belgium map") entered in the search engine: http://search.arch.be/en/zoeken-naar-archieven/zoekresultaat/index/index/zoekterm/**belgium map**/lang/en.

term(s) and the time. Time has been used here as an HTTP request identifier: it eases the distinction of several occurrences of the same query (strings are identical) during the same visit (the ID numbers are identical). Using the "group by" operation, we have been able to group rows containing the same queried term(s) and, combined with the "count method", we obtained in a single row the visit ID, the user query and the number of occurrences per visit.

4.4 Cleaning

User queries were stored in a CSV file. In order to ensure a better understanding of the queries, the decision was made to not parse queries into separate words. However, it appears that there are "hidden duplicates". In order to clean our data and unmask these similar queries, we decided to convert all characters to lowercase, to replace each character which is not alphanumeric (for example "+") by a space, to replace special characters (for example, replacing "é" by "e") and to trim whitespaces. Thus, an original user query such as "matheus+De+Vuyst" becomes "matheus de vuyst". In order to identify most requested content, the grouping operation was processed once again after this cleaning step.

4.5 Clustering

The algorithm which seems most appropriate for our context is based on the Jaro-Winkler distance (Winkler, 1990), a variant of the more common Jaro distance (Jaro, 1989). This method aims at detecting duplicates by counting the number of character substitutions needed to transform one string into another, and primarily targets names. Since user queries almost exclusively contain short keywords, the choice of this algorithm makes the most sense.

In order to make sure that the quality of the clusters was good, we only kept the results that satisfied the threshold of 0.8. The algorithm has been useful to deal with typographic errors frequently present in place names and to cluster search queries like *brusel* with their correct Dutch form, *brussel.* Perhaps more impressively, it also matched *Deinze O.L.V.* and *Onze Lieve Vrouwkerk Deinze*, two different query strings aiming to find the one and only *Onze Lieve Vrouw* ("Our Lady") church in the Belgian city of Deinze. However, it should be noted than while trying to cluster different strings, one

does induce noise and it is nearly impossible to have an algorithm that matches different-but-similar words without false positives, such as “shirt” and “t-shirt”.

5 Results

First of all, results obtained while applying our 5-step method to our case study led to a significant reduction in the data to be analysed. At each stage (table 1), we were able to use filters and different processes to keep only the chosen data and to regroup very similar queries. To sum up, we moved from an initial data set consisting of more than 13 GB of data to a final file containing about 22 000 queried terms.

Table 1: Results obtained in terms of quantity[15]

Collecting	Parsing	Grouping	Cleaning	Clustering
Filtering 30 MB of “relevant” URLs from 13 GB of raw data	Extraction of 189,000 user queries from the URLs	Skimming up to 50,000 queries by keeping each one max 1×/visit	Harmonization from 50,000 to 37,000 different queries	Obtainment of 22,000 different queries

In terms of content, it is interesting to consider how the queries with the higher frequency count are affected by the method applied before retrieving them. There is no big surprise among the three most popular queries (table 2): there are either about parish registers (“parochieregisters” in Dutch and “registres paroissiaux” in French) or civil status registers (“etat civil” in French).

Table 2: The 3 most requested terms

After Parsing	After Grouping	After Normalising	After Clustering
parochieregisters	parochieregisters	etat civil	parochieregisters
etat+civil	etat civil	parochieregisters	etat civil
etat civil	registres paroissiaux	registres paroissiaux	registres paroissiaux

15 For clarification purposes, numbers have been rounded to the nearest unit for the “collecting” step and to the nearest thousand for all other steps.

More interesting results appear if we look at the 10 most popular queries. Thus, among the queries parsed from the raw data, three of them ("liege", "tournai" and "ellezelles", which are three Belgian place names) disappear within the next steps, being apparently less representative than expected. Similarly, among the top 10 queries resulting from clustering, new entities appear, which were absolutely not present in the previous steps, such as "burgerlijke stand" (civil status, in Dutch this time), "mariage" (wedding in French) or "acte de naissance" (birth certificates in French).

Eventually, in terms of contents itself, a more in-depth analysis should be carried out in collaboration with the staff members of the institution. However, it is already possible to notice that among the most requested terms, we find some global types of archive, such as parish registers, i.e. archive that were produced in the context of a parish. This finding reveals that current access to this type of archive might not be sufficiently visible, intuitive or effective.

6 Discussion and future work

Based upon the findings presented and analysed in this paper, it appears that text mining has the potential to help cultural heritage institutions to deal with huge data sets and obtain more representative results. However, the method presented here remains relatively labor-intensive and cannot be completely automated. Elements such as URL structures are variable and require human supervision to make the necessary adaptations. Such a process seems to be worthwhile as long as time and effort devoted to that analysis are considered as an initial investment which can lead to substantial savings for the future.

While first insights, based upon a narrow selection of the most requested terms within the digital catalogue, can be useful for the staff members of the Royal Archives of Belgium, they should not detain us from also considering the other results. As underlined by Khoo et al. (2008), "websites, as Internet nodes, exhibit many of the power law distributions typical of the Internet, characterised by a small number of data with high frequency counts at one end of the distribution, and a large number of data with low frequency counts at the other end." Thus, considering only data with high frequency counts, such as the 10 most used queries, could overshadow other relevant data.

Future research will build on the method presented here in various ways. One possible way is to make comparisons in order to evaluate whether it is better to analyse queries as a whole, as performed here, or by parsing each word if they are composed of several terms. Also, more insights could be gained by identifying which words are the most often associated to some others, such as locations accompanying parish registers. Moreover, it would probably lead to more significant and interesting data to use a list of stop words including determinants and prepositions. Another research avenue could be to focus on visualising the evolution of queries in time. Eventually, one could resort to using named-entity recognition, which aims at detecting interesting entities in text and assign them a type, such as person, location or product, and therefore provides insights as to what kind of queries the visitors are interested in.

Acknowledgements

The authors would like to extend their gratitude to Raphaël Hubain, Seth van Hooland, Florence Gillet, Martin Vanbrabant, Gerald Haesendonck, Jill Hungenaert, Emmylou Haffner and Mathias Coeckelbergs.
The research underlying the results presented in this article was funded by the Belgian Science Policy Office in the context of contract number BR/153/A5/MADDLAIN.

References

Ceccarelli, Diego, Sergiu Gordea, Claudio Lucchese, Franco Maria Nardini, and Gabriele Tolomei (2011): Improving Europeana Search Experience Using Query Logs. In: *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2011* (pp. 384–395). Berlin, Heidelberg: Springer.

Chase, Darren, Elizabeth Trapasso, and Robert Tolliver (2016): The Perfect Storm: Examining User Experience and Conducting a Usability Test to Investigate a Disruptive Academic Library Web Site Redevelopment. In: *Journal of Web Librarianship*, 10 (1), 28–44. doi:10.1080/19322909.2015.1124740

Dijkshoorn, Chris, Lora Aroyo, Guus Schreiber, Jan Wielemaker, and Lizzy Jongma (2014): Using Linked Data to Diversify Search Results a Case Study in Cultural Heritage. In: K. Janowicz et al. (Eds.): *Knowledge Engineering and Knowledge*

Management. 19th International Conference, EKAW 2014, Linköping, Sweden, Proceedings (pp. 109–120).

Goyvaerts, Jan, and Steven Levithan (2012): *Regular Expressions Cookbook*. O'Reilly Media.

Grimes, Carrie, Diane Tang, and Daniel M. Russell (2007): Query Logs Alone are not Enough. In: *Proceedings of the Workshop on Query Log Analysis: Social and Technology Challenges at the 16th International World Wide Web Conference (WWW 2007), Banff, Canada, May 8–12, 2007*. http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/34431.pdf

Jaro, M. A. (1989): Advances in Record Linkage Methodology as Applied to the 1989 census of Tampa Florida. In: *Journal of the American Statistical Society*, 84 (406), 414–420.

Kathuria, Ashish, Bernard J. Jansen, Carolyn Hafernik, and Amanda Spink (2010): Classifying the User Intent of Web Queries using K-means Clustering. In: *Internet Research*, 20 (5), 563–581.

Khoo, Michael, Joe Pagano, Anne L. Washington, Mimi Recker, Bart Palmer, and Robert A. Donahue (2008): Using Web Metrics to Analyze Digital Libraries. In: *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL '08), Pittsburgh, USA, June 16–20, 2008*, New York (pp. 375–384).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2009): *Introduction to Informational Retrieval* (online edition). Cambridge University Press. http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf

Showers, Ben (2015): *Library, Analytics and Metrics: Using data to drive decisions and services*. Facet Publishing.

van Hooland, Seth, Françoise Vandooren, and Eva Méndez Rodríguez (2011): Opportunities and Risks for Libraries in Applying for European Funding. In: *The Electronic Library*, 29 (1), 90–104.

Winkler, William E. (1990): String Comparitor Metrics and Enhanced Decision Roles in the Fellegi-Sunter Model of Record Linkage. In: *Proceedings of the Section on Survey Research Methods*. Washington, DC: American Statistical Association (pp. 354–359).

Woods, Roberta (2010): From Federated Search to the Universal Search Solution. In: *The Serials Librarian*, 58 (1), 141–148.

Zavalina, Oksana (2007): Collection-Level User Searches in Federated Digital Resource Environment. In: Andrew Grove (Ed.*): Joining research and practice: social computing and information science / ASIS&T 2007, October 19–24, Milwaukee, Wisc.* Proceedings of the ASIS&T Annual Meeting; Vol. 44, Issue 1. Silver Spring, Md.: American Society for Information Science and Technology. doi:10.1002/meet.1450440225

Zavalina, Oksana, and Elena V. Vassilieva (2014): Understanding the Information Needs of Large-Scale Library Users. In: *Library Resources & Technical Services*, 58 (2), 84–99.